Basic Univariate Statistics

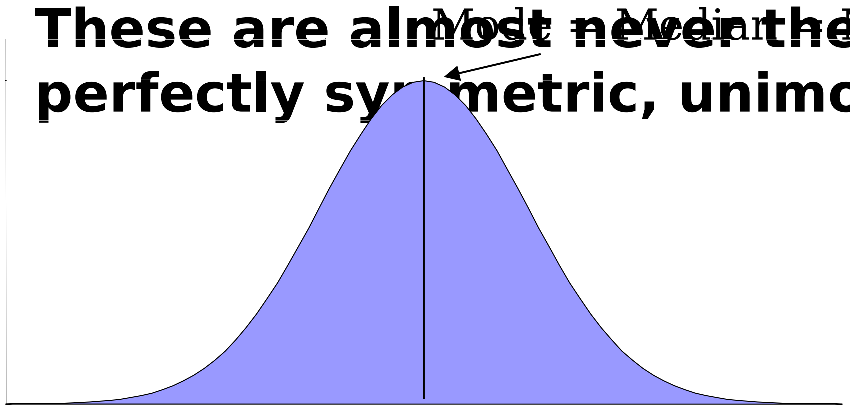
Chapter 6

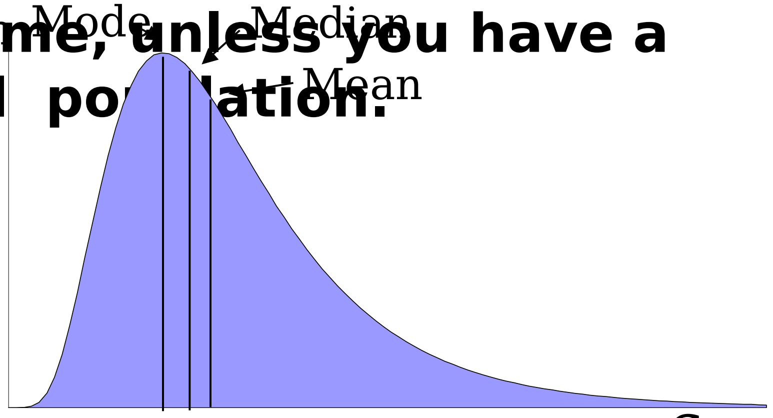
Basic Statistics

- **A statistic is a number, computed from sample data, such as a mean or variance.**
- **The objective of Statistics is to make *inferences* (predictions, decisions) about a *population* based upon information contained in a *sample*.**
  - » **Mendenhall**
- **The objective of Statistics is to make *predictions* about the cost of a *weapon system* based upon information in *analogous systems.***
  - » **DoD Cost Analyst**

Measures of Central Tendency

- These statistics describe the "middle region" of the sample.
  - Mean
    - » The arithmetic average of the data set.
  - Median
    - » The "middle" of the data set.
  - Mode
    - » The value in the data set that occurs most frequently.
- These are almost never the same, unless you have a perfectly symmetric, unimodal population.





Mean

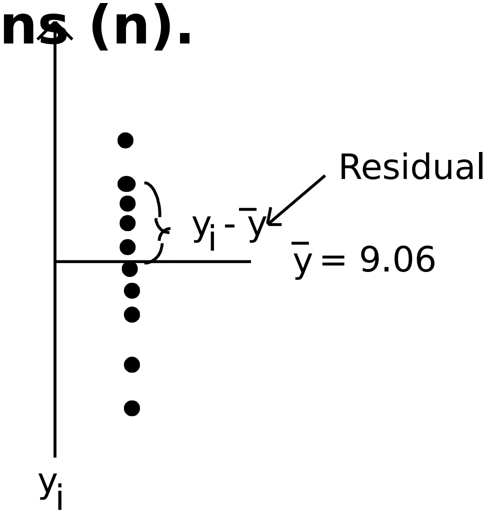
- **The Sample Mean ($\bar{y}$) is the arithmetic average of a data set.**
- **It is used to estimate the population mean, (μ).**
- **Calculated by taking the sum of the observed values (y_i) divided by the number of observations (n).**

Historical Transmogrifier Average Unit Production Costs

System	FY97$K
1	22.2
2	17.3
3	11.8
4	9.6
5	8.8
6	7.6
7	6.8
8	3.2
9	1.7
10	1.6

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

$$\bar{y} = \frac{22.2 + 17.3 + \cdots + 1.6}{10} = \$9.06K$$



Median

- The Median is the middle observation of an ordered (from low to high) data set
- Examples:
  - 1, 2, 4, 5, 5, 6, 8
    - » Here, the middle observation is 5, so the median is 5
  - 1, 3, 4, 4, 5, 7, 8, 8
    - » Here, there is no "middle" observation so we take the average of the two observations at the center

$$\text{Median} = \frac{4+5}{2} = 4.5$$

Mode

- **The Mode is the value of the data set that occurs most frequently**
- **Example:**
  - **1, 2, 4, 5, 5, 6, 8**
    - » **Here the Mode is 5, since 5 occurred twice and no other value occurred more than once**
- **Data sets can have more than one mode, while the mean and median have one unique value**
- **Data sets can also have NO mode, for example:**
  - **1, 3, 5, 6, 7, 8, 9**
    - » **Here, no value occurs more frequently than any other, therefore no mode exists**
    - » **You could also argue that this data set contains 7 modes since each value occurs as frequently as every other**

Dispersion Statistics

- The Mean, Median and Mode by themselves are not sufficient descriptors of a data set
- Example:
  - Data Set 1: 48, 49, 50, 51, 52
  - Data Set 2: 5, 15, 50, 80, 100
- Note that the Mean and Median for both data sets are identical, but the data sets are glaringly different!
- The difference is in the dispersion of the data points
- Dispersion Statistics we will discuss are:
  - Range
  - Variance
  - Standard Deviation

Range

- **The Range is simply the difference between the smallest and largest observation in a data set**
- **Example**
  - **Data Set 1: 48, 49, 50, 51, 52**
  - **Data Set 2: 5, 15, 50, 80, 100**
- **The Range of data set 1 is 52 - 48 = 4**
- **The Range of data set 2 is 100 - 5 = 95**
- **So, while both data sets have the same mean and median, the dispersion of the data, as depicted by the range, is much smaller in data set 1**

Variance

- **The Variance, s^2, represents the amount of variability of the data relative to their mean**
- **As shown below, the variance is the “average” of the squared deviations of the observations about their mean**

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

- **The Variance, s^2, is the *sample* variance, and is used to estimate the actual *population* variance, σ^2**

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N}$$

Standard Deviation

- **The Variance is not a "common sense" statistic because it describes the data in terms of squared units**
- **The Standard Deviation, s, is simply the square root of the variance**

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

- **The Standard Deviation, s, is the *sample* standard deviation, and is used to estimate the actual *population* standard deviation, σ**

$$\sigma = \sqrt{\frac{\sum (y_i - \mu)^2}{N}}$$

Standard Deviation

- **The sample standard deviation, s, is measured in the same units as the data from which the standard deviation is being calculated**

System	FY97$K	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	22.2	13.1	172.7
2	17.3	8.2	67.9
3	11.8	2.7	7.5
4	9.6	0.5	0.3
5	8.8	-0.3	0.1
6	7.6	-1.5	2.1
7	6.8	-2.3	5.1
8	3.2	-5.9	34.3
9	1.7	-7.4	54.2
10	1.6	-7.5	55.7
Average	**9.06**		

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$= \frac{172.7 + 67.9 + \cdots + 55.7}{10-1}$$

$$= \frac{399.8}{9} = 44.4(\$K^2)$$

$$s = \sqrt{s^2} = \sqrt{44.4(\$K^2)}$$

$$= 6.67(\$K)$$

- **This number, $6.67K, represents the *average estimating error for predicting subsequent observations***
- **In other words: *On average, when estimating the cost of transmogrifiers that belongs to the same population as the ten systems above, we would expect to be off by $6.67K***

Coefficient of Variation

- **For a given data set, the standard deviation is $100,000.**
- **Is that good or bad? It depends…**
  - **A standard deviation of $100K for a task estimated at $5M would be very good indeed.**
  - **A standard deviation of $100K for a task estimated at $100K is clearly useless.**
- **What constitutes a "good" standard deviation?**
- **The "goodness" of the standard deviation is not its value per se, but rather what percentage the standard deviation is of the estimated value.**
- **The *Coefficient of Variation* (CV) is defined as the "average" percent estimating error when predicting subsequent observations within the representative population.**
- **The CV is the ratio of the standard deviation to the mean.**

$$CV = \frac{s_y}{\bar{y}}$$

Coefficient of Variation

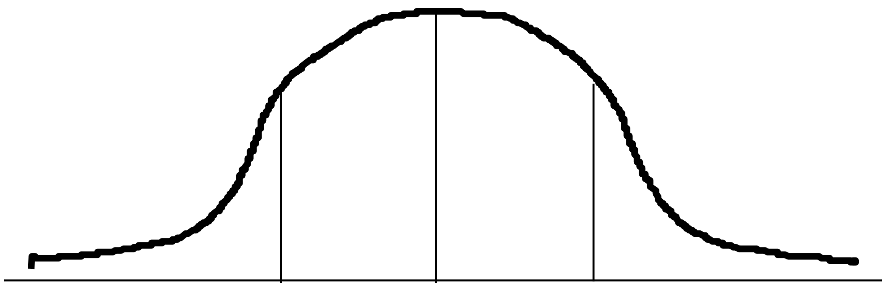
- **In the first example, the CV is $100K/$5M = 2%**
- **In the second example, the CV is $100K/$100K = 100%**
- **These values are unitless and can be readily compared.**
- ***The CV is the "average" percent estimating error for the population when using $\bar{y}$ as the estimator.***
- **Or, *the CV is the "average" percent estimating error when estimating the cost of future tasks.***
- **Calculate the CV from our previous transmogrifier cost database:**
  - **CV = $6.67K/$9.06K = 73.6%**
- **Therefore, *for subsequent observations we would expect to be off on "average" by 73.6% when using $9.06K as the estimated cost*.**

Normal Distribution

•Continuous random variables are associated with populations which contain an infinitely large number of observations

• The heights of trees, the weight of humans, the cost of weapons, and the error when predicting the cost of a weapon system are all examples of continuous random variables.

• These continuous random variables can take on a variety of shapes, many of which approximate a bell-shaped curve. These bell-shaped curves approximate the normal probability distribution.



The Empirical Rule: Given a distribution of measurements drawn from a population that is approximately bell-shaped, the interval

1. will contain approximately 68 percent of the measurements.
2. will contain approximately 95 percent of the measurements.
3. will contain approximately all (99.7%) of the measurements.